IDA

INSTITUTE FOR DEFENSE ANALYSES

Analysis of VSAM Research at Carnegie Mellon University and the Sarnoff Corporation: Potential Application to Small Unit Operations

Cynthia Dion-Schwarz

December 1999

Approved for public release; distribution unlimited.

IDA Paper P-3428

Log: H 98-002403

20001212 005

This work was conducted for the Defense Advanced Research Projects Agency, Project Assignment A-210. The publication of this IDA document does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper P-3428

Analysis of VSAM Research at Carnegie Mellon University and the Sarnoff Corporation: Potential Application to Small Unit Operations

Cynthia Dion-Schwarz

PREFACE

This work was completed under a task entitled "Tactical Technical Analyses," which tasked the Institute for Defense Analyses (IDA) to support the Defense Advanced Research Projects Agency's (DARPA's) program called Small Unit Operations/ Situational Awareness System (SUO/SAS). IDA assisted DARPA in assessing current technologies that could be used with small units to enhance situational awareness. IDA was also asked to provide assessments of the results of field experimentation of various sensors used with small units. The DARPA program manager of this task was Dr. Mark McHenry, Program Manager SUO/SAS, DARPA.

CONTENTS

FIGURES

TABLES

I. INTRODUCTION

This paper summarizes an analysis of the Video Surveillance and Monitoring (VSAM) research that is being conducted by researchers at Carnegie Mellon University (CMU) and the Sarnoff Corporation under the Defense Advanced Research Projects Agency (DARPA) Image Understanding Program. The researchers are studying ways to create a VSAM system for battlefield management. However, this system also has clear applications to Small Unit Operations (SUO) and should be developed and adapted for this program. The analysis and comments that follow summarize the present research and development (R&D) status of the VSAM program that would have applications in SUO.

An analysis of the motion-detection algorithms developed by the researchers shows that this system can accurately detect and track motion from humans, human groups, and vehicles. This system tracks best when sufficient numbers of pixels are on target (e.g., when a 2-m human is equivalently 300 m or less in range, giving at least 4×8 pixels on target). The detection probability is nearly 100 percent, and the false-alarm rate is probably less than 0.22 per minute. Further study reveals that these false alarms are often caused by tracking errors rather than by detection errors when tracking is done in real time. This means that these false alarms can occur primarily when a true target is present in the field of view, thus mitigating the effect of a false alarm.

The most encouraging aspect of these results, however, is that these high detection rates and relatively low false-alarm rates were derived from a single video surveillance camera. As conceived, VSAM is a *system* of integrated video surveillance units, complete with sensor-to-sensor analysis and site modeling. Such a system would reduce the false-alarm rates to a negligible level, depending on the image overlap and the quality of the site model. The results herein should be regarded as a lower limit to the capabilities of the VSAM technologies being developed.

II. A DESCRIPTION OF VSAM RESEARCH GOALS AND STATUS

VSAM research is being conducted by a collaboration of academia and industry (the Robotics Institute at CMU and the Sarnoff Corporation) under the DARPA Image Understanding Program.[1] The researchers' goal is to develop a cooperative, multi-sensor video surveillance system for large battlefield areas. To achieve these goals, the team is developing software and integrating ordinary, inexpensive commercial-off-the-shelf (COTS) hardware systems that will automatically track and identify moving targets, such as soldiers, groups of soldiers, and vehicles. The team is concentrating on integrating the hardware and developing the coordinated tracking algorithms (software) that provide good target identification (ID) and target tracking with a low rate of false alarms.

Much of the VSAM technology that the CMU-Sarnoff research team is developing can directly benefit the SUO project. Both projects[2] need to identify and track multiple targets that pose a potential threat in sometimes confusing environments, monitor environments for unusual events and activities, derive an automated system for activities analysis, and obtain extended coverage of an area by using networks of sensors. Both projects also need to maintain a low rate of false alarms.

Other requirements are less similar. For instance, the distance scales over which the VSAM project may need to monitor are likely larger than what would be needed in the SUO program. In addition, the VSAM project relies on a good site model to range targets accurately, whereas video equipment deployed for SUO would likely be placed in less-well-surveyed areas. Lastly, for the small-scale, dynamic situations envisioned for the SUO project, the resources (such as battery power, communications bandwidth, and weight) are much more limited than the resources that would be encountered in a large-scale battlefield. Nevertheless, this paper will show that the VSAM project has applications that would benefit SUO.

---

1 DARPA Broad Agency Announcement (BAA) 96-14.

2 The VSAM project and the SUO project

The VSAM technologies developed by this team typically include several charged-coupled device (CCD) cameras mounted on pan-tilt hardware that is controlled by a Pentium personal computer (PC). The images are analyzed for motion and classification. A central control unit that has communication links with all the remote PC/camera units handles coordination of tracking among sensors. Other imaging devices, such as infrared (IR) sensors, stereo sensors, or laser-ranging sensors, can also be used. Novel omni-directional cameras with automated detection and tracking show great promise, particularly in restricted terrain (e.g., wooded or urban areas).

A PC associated with each camera processes the images, locates targets, and controls the camera's tracking of the target by using a combination of pans, tilts, and zooms. In addition, researchers plan to have the PC multi-task multiple targets (i.e., the camera's pan-tilt, real-time tracking switches among targets if several are in the field of view). For real-time monitoring, the data sent from the PC/camera system to the central control unit could be minimized to include only the tracking/target ID information (low bandwidth), rather than high-bandwidth images. Coordination (i.e., hand-off of images from one system to another) is done by the central control unit and, again, includes minimal data transfer to a PC/camera system. The only high-data transfer that might occur is when the operator of the central control unit wants to look at the actual images produced by a remote camera (which might be desirable for visual confirmation of a target). Provided that false-alarm rates are minimal-to-nonexistent, such a system would provide a low-bandwidth video monitor system that could be deployed in a variety of situations encountered by SUO.

The PC/CCD camera systems could be relatively lightweight and low powered, and communications could be handled with low-power radio systems. Such a system could also be made rugged enough for a variety of weather systems and terrain. Lastly, since the hardware components are relatively inexpensive and commercially available, multiple units could be built and deployed without large incremental costs.

A group of such sensors could be deployed in several locations that small units might want to monitor. For example, a group of soldiers may wish to monitor the perimeter of a secured area. These sensors could be deployed around the perimeter—in communication with a central control unit—and give the soldiers automatic notification about incursions. As another example, in an urban area, the entrances to a building or street may need to be monitored before the small unit secures the area. An advance party could

drop a few sensors into the area, and the system could be used to analyze activities automatically and identify threatening behavior (e.g., rioting, loitering, and so forth).

Our preliminary analysis shows that these systems may be useful in the night (using IR), in the day, and in a variety of weather patterns. A single human operator could monitor many sensors at one time. Thus, this system, with some modifications and adaptations, would be useful in the different situations encountered by small units.

III. HOW THE ALGORITHMS WORK

The results presented here evaluate the tracking algorithms developed by the CMU-Sarnoff research team to detect motion from video. A single PC/video camera system was analyzed. There are two types of algorithms under development. Both algorithms use the same method to detect the actual motion; however, they differ in the way in which the background to the motion is calculated.

In the first algorithm, called motion detection by *temporal differencing*, the background to the motion is simply a previous frame. Motion is evaluated by subtracting the current frame from a previous frame separated by a fixed period of time and searching for changes in the gray-scale levels, proximity, and numbers of pixels. The amount of gray-scale change required to call the changes movement can be adjusted as a threshold by the operator. This would allow the observer to adjust the criteria for movement as the contrast in the scene changes.

In the second algorithm, called motion detection by *adaptive background subtraction*, the background is modeled by calculating a running average of the gray-scale value of each pixel in all the previous frames. In a scene with movement, the value assigned to a pixel "jumps" to a new gray-scale value. If that new background is consistent with movement (i.e., other neighboring pixels "jump" to new values and continue to change dramatically with each frame), the target is tracked. Otherwise, the algorithm identifies the pixel change with a new background value and starts developing a new model for that pixel. As a result, a target that is moving is identified, but, if it later stops, it becomes a part of the background until it begins moving again. Moving targets, once identified, can be tracked even if stationary, but this feature was not enabled during this analysis nor evaluated for this paper.

Both algorithms identify targets by looking for proximate pixel changes. Many pixels grouped together forming an outline can be identified depending on their relative height-to-width ratios. That is, wide objects that move are "vehicles," and narrow objects that move are "human." As a result, using this method, human groups (which can be rather wide compared with their height) are frequently misidentified as vehicles. However, researchers are presently developing a new target ID algorithm that appears to

distinguish vehicles from human groups correctly. Some preliminary results about this new algorithm are also presented here.

One important feature that determines how well these algorithms detect and track motion is the numbers of pixels on target. For the CMU-Sarnoff researchers' algorithms to detect and track the motion, 8 × 4 pixels—at a minimum—should be on target using their modest-resolution 320 × 240-pixel CCD cameras. The Institute for Defense Analyses (IDA) cameras that were used to film the scenes were higher resolution cameras (640 × 480 pixels). In the video examples described in this paper, a 2-m human moving 600 m from the (unzoomed) camera spanned approximately 8 × 4 pixels and was thus barely detectable by the motion-detection algorithms. If CMU's cameras were zoomed (which changes the focal length of the camera and the number of pixels on target but also reduces the field of view) or if different, higher resolution cameras were used, the human motion detection by this system is changed. That is, humans moving at 600 m from the camera but viewed with a ×2 zoom have twice as many pixels on target and, thus, are detected much more easily.

In this paper, up to three different cameras of identical resolution were used. All the cameras had the same true range to the movement (approximately 600 m), but the zoom settings were different to test the ability of the camera to detect motion depending on the number of pixels on target. The three possible settings were:

1. The wide-view setting (approximately a 60-deg field of view with a ×1 zoom)
2. The medium-view setting (approximately a 41-deg field of view with a ×2 zoom)
3. The narrow-view setting (approximately a 23-deg field of view with a ×4 zoom).

IV. WHAT DATA WERE AVAILABLE AND WHAT WERE THE LIMITATIONS OF THESE DATA?

Two types of data were analyzed. The first type was a set of two edited Hi-8 video tapes (about 2 hours total) that IDA researchers filmed at Fort Benning, Georgia, in spring 1998. The original Hi-8 films included scenes of troop and tank exercises. Three cameras set at three different zooms filmed two scenes of troop exercises. Two additional scenes of troop exercises and one scene of tank maneuvers were filmed by two cameras set at two different zooms. The film footage was later edited, copied onto Hi-8 tape, and taken to CMU for motion (tracking) analysis.

The CMU Robotics group also collected a second type of data during the filming at Fort Benning in April. A CCD camera filmed similar scenes, and the movements of troops and tanks were tracked in real-time. The resulting recorded film contained both the original video of the scene and the tracks overlaid by the computer program. This video was transferred to Hi-8 video and analyzed further.

Thus, one set of data comprises the edited, copied film that was later analyzed for motion and other parameters. The second set of data comprises the result of a real-time tracking of motion. Because all the IDA film was edited and later transferred to another tape before the motion analysis was done, the tapes were probably noisier than those obtained with the "live" real-time scenes provided by CMU. This may result in a higher false-alarm rate than one would obtain in real-time. Nevertheless, the results from both analyses probably give a fair lower limit of how this system would perform in a variety of situations encountered by SUOs.

V. ANALYSIS METHOD

In the IDA-filmed troop exercises, the scenes comprised individuals or groups conducting exercises in and near a stand of trees. The weather was sunny, but fairly windy, leading to some motion-detection false alarms. The movements of the troops varied: close to the camera to far from the camera; in the trees (so that soldiers were sometimes obstructed); and running, walking, and bounding across the camera's field of view. Occasionally, soldiers loomed or receded to and from the camera's field of view. These views were filmed with the three cameras, so three different zooms (numbers of pixels on target) were available of the same scenes. This enabled the evaluation of the program's motion detection ability as a function of numbers of pixels on target.

In the IDA-filmed tank and Armored Personnel Carrier (APC) maneuvers, the tanks/APCs were filmed with two different zooms: one with no zoom and one with a ×2 zoom and both at a range of 400–600 m). Drizzle was falling, and the sky was dark and cloudy. The tanks/APCs were filmed as they traveled back and forth along a tree line. In the film with no zoom, detecting the tank/APC movement—even by a human observer—was difficult. This same vista also filmed with three soldiers walking near the tree line was also difficult to detect.

These tapes were taken to CMU and analyzed for motion by IDA researchers. The tapes were played in a camcorder, and the video was fed into a monitor and two computers running two different analysis programs simultaneously. The computers displayed both the video feed and overlaid graphics indicating the detection of motion (typically, a box drawn around the moving target). This target was classified as "human" or "vehicle" by overlaid screen graphics.

The results were analyzed by comparing the video footage by eye with the results displayed by the computer. Data such as the amount of time to target ID were estimated by using the camcorder's clock. Other data, such as the false-alarm rate and persistence of false alarms, were collected. Lastly, those targets that were clearly visible but that the computer failed to detect were noted. Some of the data were evaluated using both the temporal differencing method of background evaluation and the adaptive background subtraction algorithm (see Section III). Most of the data were evaluated using the

temporal differencing method of background evaluation because the CMU researchers were more confident about its robustness.

Because of time constraints, all the video footage could not be analyzed (i.e., collecting timing data). All the video was studied, however, and the samples presented here are representative of the total results. In addition, gray-scale thresholds that define the movement were adjusted to study the detection/false-alarm rates as a function of these thresholds. Section VII contains the results.

The second set of data provided by CMU comprised a 10-minute tape of the real-time analysis of the motion. The algorithm that CMU researchers used was the temporal differencing method of background estimation. The scenes comprised troop movements similar to the ones filmed by IDA but, typically, at a single, narrow (zoomed) field of view. Again, the weather varied (windy and sunny, rainy and dark). These tapes were analyzed at IDA for numbers and persistence of false alarms and for the detection rates of troops. These rates were again estimated by eye and using a clock. Section VII also contains the results of this analysis. Figures 1 and 2 show stills from the video provided by CMU. In Figure 1, three human targets are detected. The soldiers are outlined with rectangles and identified correctly. In Figure 2, a vehicle (tank) is detected and again identified correctly.

A third set of data not taken at Ft. Benning but studied by IDA researchers is worth mentioning here. These data comprised live video of a busy CMU campus parking lot that was analyzed in real time. The method used by CMU researchers is under development, using the adaptive background subtraction technique and better target ID. In this video, human groups were distinguished from vehicles through shape recognition. Section VI summarizes the results.

The evaluations of the VSAM algorithm relied on using the "human eye" as a benchmark for performance. This works well, but certain quantities could be provided to make the analysis easier. In particular, as the tracking algorithm analyzes data, it should create a tracking report that records information as it is collected. For instance, with each target acquired, the frame number (equivalent to elapsed time), a target number (to differentiate targets), mean pixel location (to locate the target approximately in the view), and target ID could be recorded. From these data, the analyst can derive target duration and locations. This information could then be compared to what is actually viewed on the screen and enable the analyst to identify true targets and false alarms quickly and to acquire statistics.

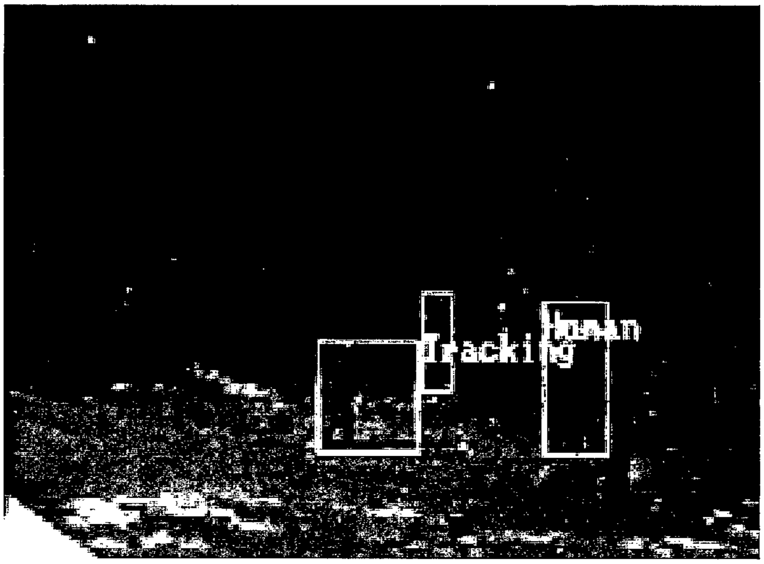

Figure 1. Still of Human Targets Detected by the VSAM Automatic Motion Detection Algorithm

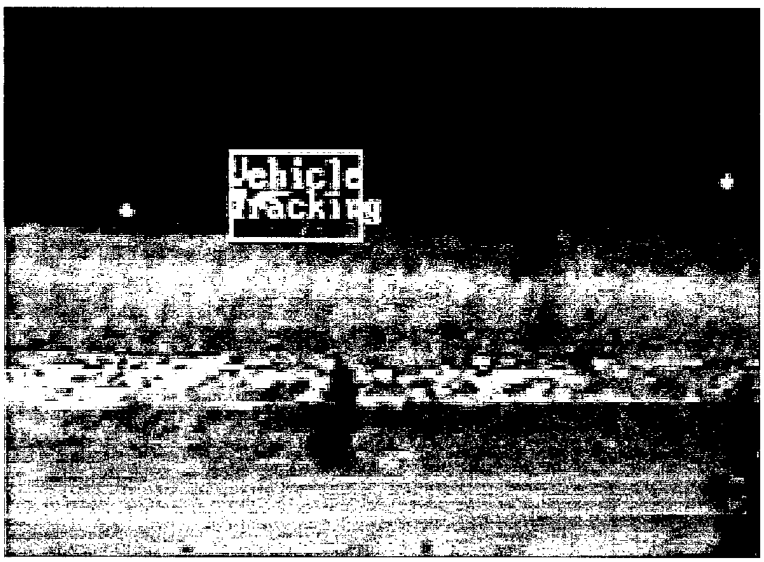

Figure 2. Still of a Vehicle Detected by the VSAM Automatic Motion Detection Algorithm

VI. RAW DATA COLLECTED

Several action sequences were analyzed for target acquisition and false-alarm rates quantitatively. All scenes were evaluated qualitatively. Table 1 summarizes the data analyzed and the type of analysis performed.

Table 1. A Description of the Data Presented in This Paper

Scene Number	Source of Data	Type	Zoom Setting	Scene Content/Range	Weather Conditions	Quantitative Analysis (Numbers)
1	IDA	Wide-view video tape	×1	Soldiers conducting exercises/~200 m	Sunny and windy	Yes
2	IDA	Medium-view video	×2	Same as Scene 1	Sunny and windy	Yes
3	IDA	Narrow-view video	×4	Same as Scene 1	Sunny and windy	Yes
4	IDA	Wide-view video	×1	Soldiers conducting exercises/~200 m	Sunny and windy	No
5	IDA	Medium-view video	×2	Same as Scene 4	Sunny and windy	No
6	IDA	Narrow-view video	×1	Same as Scene 4	Sunny and windy	No
7	IDA	Wide-view video tape	×1	Tanks conducting exercises/~600 m	Overcast and raining	Yes
8	IDA	Medium-view video tape	×2	Same as Scene 7	Overcast and raining	Yes
9	IDA	Wide-view video tape	×1	Tanks conducting exercises (~ 600 m)	Overcast and raining	Yes
10	IDA	Medium-view video tape	×2	Same as Scene 9	Overcast and raining	Yes
11	IDA	Wide-view video tape	×1	Soldiers walking in tree line (~ 600 m)	Overcast and raining	Yes
12	IDA	Medium-view video tape	×1	Same as scene 11	Overcast and raining	Yes
13	CMU	Narrow-view film of real-time analyzed scenes	×4	Different views of soldiers conducting exercises	Sunny and windy	Yes
14	CMU	Live, real-time video feed	×1	Busy parking lot at CMU	Sunny and calm	No

Scene 1 (wide-view scene of soldiers conducting exercises) was analyzed using both the temporal differencing method and adaptive subtraction method of background averaging. The gray-pixel threshold of both algorithms was set to the same level (T = 35).[3] A total of 16 possible moving targets was contained in the 442 seconds of film. The scene was quite windy and sunny. Table 2 shows the results.

Table 2. Raw Results From the Analysis of Scene 1: a Wide (Far) View of Soldier Exercises

Target Number	Comments	Temporal Differencing Algorithm Time To Acquire Target (sec)	Adaptive Background Subtraction Algorithm Time To Acquire Target (sec)
1	Soldier looming	32	6
2	Soldier far away	Not acquired	0
3	Close to soldier	11	0
4	L to R walking soldier	31	Not acquired
5	Looming soldier	6	3
6	Looming soldier	0	0
7	Far away soldier	Not acquired	2
8	Far and obstructed soldier	Not acquired	1
9	L to R and far away soldier	Not acquired	Not acquired
10	Looming soldier	5	0
11	L to R soldier	0	0
12	R to L soldier	0	1
13	Far away soldier	Not acquired	0
14	Looming solider	Not acquired	0
15	L to R soldier	Not acquired	2
16	Looming and running soldier	0	1

Soldiers who loom (walk toward the camera) are typically difficult to detect, as can be seen by the long times taken to acquire (if at all) the looming soldiers. The temporal differencing algorithm acquired and tracked 9 of the potential 16 moving targets, and the adaptive background subtraction algorithm acquired 14 moving targets. Only one target (number 9) was never detected by either algorithm.

---

3 The threshold of T = 35 refers to the gray level at which the pixel is considered to be "in motion." Gray-scale levels used in this analysis were from 0–255 (8-bits).

Although the adaptive background subtraction algorithm had a higher rate of detection, it also had a higher false-alarm rate. A total of 19 false alarms was detected by the adaptive background subtraction algorithm in the 442 seconds of film, and the computer dwelled and tracked these false alarms for a total of 110 seconds. Thus, 25 percent of the time in which a target was detected was spent on false alarms. The temporal differencing algorithm appeared to be more robust, detecting no false alarms.

Scene 2 (medium-view film that contained the same images as Scene 1) was also analyzed using both the temporal differencing and adaptive background subtraction methods of background averaging. The gray-pixel threshold of both algorithms was unchanged (T = 35). A total of 13 possible moving targets was contained in 445 seconds of film. Fewer images appear because some soldiers were cut off in this more narrow view, and the tape ran slightly longer because it was filmed using a different camera. However, again, the scene was quite windy and sunny. Table 3 shows the results.

Table 3. Raw Results From the Analysis of Scene 2: a Medium View of Soldier Exercises

Target Number	Comments	Temporal Differencing Algorithm Time To Acquire Target (sec)	Adaptive Background Subtraction Algorithm Time To Acquire Target (sec)
1	Looming soldier	15	8
2	3 men walking	9	45
3	1 man L to R	2	3
4	1 man L to R	4	2
5	1 man running and looming	Not acquired	7
6	Far, 2 men	Not acquired	3
7	None	0	5
8	L to R soldier	1	1
9	3 men running	1	Not acquired
10	L to R and far	1	3
11	1 man running	1	5
12	R to L soldier	0	0
13	3 men looming	1	2

In this medium view, both algorithms show little difference in their ability to acquire and track targets. Overall, however, the temporal differencing algorithm appears to acquire targets more quickly, although possibly less frequently (this is only "possibly" because of the small numbers of possible targets).

Again, however, the adaptive background subtraction algorithm had a higher false-alarm rate. A total of 12 false alarms was detected by the adaptive background subtraction algorithm in the 445 seconds of film, and the computer dwelled on and tracked these false alarms for a total of 32 seconds. Thus, although less time (7 percent) was spent on false alarms, there were nearly as many false alarms as real targets. The temporal differencing algorithm again appeared to be more robust, detecting only a single false alarm for about 5 seconds (1 percent of the time). The false alarms detected by the adaptive background subtraction algorithm appeared to be caused mostly by the wind. For instance, trees swaying in the wind or flags flapping in the wind may register a false alarm. However, this adaptive background subtraction algorithm is still under development, and the CMU researchers did not expect it to be as robust as the temporal differencing algorithm.

Given that the temporal differencing algorithm did appear to be more robust, the rest of the analysis was performed using only this one. Scene 3 comprised the same vista as that filmed in Scenes 1 and 2, but this time with the most narrow view. These data were analyzed using the temporal differencing algorithm and also by adjusting the threshold value to compare the sensitivity of the algorithm to movement and false alarms. Table 4 summarizes the raw results found in this view. More targets appear than were previously noted because targets that start and stop are counted twice (i.e., the targets here are not necessarily distinct). The algorithm has the capability to lock-on to a target once it is acquired, but this feature was not enabled during this test. This provided more statistics for analysis since a restarted target is treated independently.

Table 4. Thresholds Tested on the Analysis of Scene 3

	Threshold 1	Threshold 2	Threshold 3	Threshold 4
Absolute value	25	30	35	40
Percentage change in gray scale	9.8	11.88	13.78	15.78

Four different thresholds were used to test the probability of detection versus false-alarm rate for the various thresholds. The threshold is the gray-scale value that the pixel must change before the pixel is tagged as potentially part of a moving target. The gray-scale used here was 0–255, so a threshold of 25 means that the pixel must lighten or darken approximately 10 percent before it is associated with a moving target. The threshold chosen depends on the overall contrast in the scene and somewhat on background motion. Table 4 shows the thresholds tested here. Using these thresholds, Table 5 shows the raw information that was collected.

Table 5. Raw Results From the Analysis of Scene 3: a Narrow (Close) View of Soldier Exercises Using Four Gray-Scale Thresholds for Detection of Motion

Moving Target Number	True Time In View (sec)	Threshold 1, Time On Target (sec)	Threshold 2, Time On Target (sec)	Threshold 3, Time On Target (sec)	Threshold 4, Time On Target (sec)
1	60	60	60	50	40
2	50	0	0	0	0
3	20	20	20	20	10
4	80	80	80	80	60
5	10	0	0	0	10
6	60	60	60	60	50
7	60	40	40	40	40
8	10	10	10	10	10
9	20	20	20	20	20
10	70	70	70	70	70
11	70	40	70	70	50
12	10	0	0	0	0
13	10	0	0	0	0
14	30	30	30	30	30
15	10	10	10	10	0
16	10	10	10	10	10
17	10	10	10	10	10
18	10	0	10	10	10
19	10	0	10	10	0
20	10	0	10	10	0
21	50	0	30	30	50
Totals	**670**	**460**	**550**	**540**	**470**

The total time spent on moving targets is more than the total duration of the video because sometimes several targets were in the field of view at once. Several targets (numbers 2, 12, and 13) were partially obstructed by other targets because they were a part of human groups. A human observer could identify those targets as distinct, but the computer algorithm could not always distinguish two or more targets until they were sufficiently separated. Threshold numbers 2 and 3 could observe all those targets that were clearly separated.

Two detection probabilities can be computed:

1. The probability that every individual will be detected, which depends on their movement and separation
2. The probability that every moving human and human group will be detected, which depends only on movement.

Table 6 summarizes some results of the analysis from Scene 3. These results are discussed in the following section.

Table 6. Summary of Further Results From the Analysis of Scene 3

Parameters	Units	Truth	T 1	T 2	T 3	T 4
Total time spent on targets	Seconds	670	460	540	550	460
Percentage spent on targets	%	100	69	81	82	69
Mean time spent on targets	Seconds	39.4	30.7	31.8	32.4	32.9
Median time spent on targets	Seconds	20	10	10	10	10
Total Number of human targets detected	Seconds	21	15	17	17	14
Number of human and human group targets detected	–	17	15	17	17	14
Efficiency of detecting human targets	%	–	71	81	81	67
Efficiency of detecting human/human group targets	%	–	88	100	100	82
Number of false alarms	–	0	0	2	8	6
Total time spent on false alarms	–	0	0	20	100	80
Number of false alarms/Number of detected targets	%	–	0.0	3.7	18.2	17.4
False-alarm time/time spent on targets	%	–	0	12	47	43
Mean time spent on false alarms	Seconds	–	0	10.0	12.5	13.3

Scenes 4, 5, and 6 were not analyzed quantitatively. These scenes comprised soldiers conducting exercises with similar vistas and movements found in Scenes 1–3. As found earlier, the temporal differencing algorithm appeared to have fewer false alarms. The two that were recorded in the wide view were caused by wind. Looming targets appear to take longer to be acquired, but the algorithm eventually identified most targets. In the medium view of the same scene, no false alarms were recorded, and all targets were identified. In the narrow view, one false alarm was recorded, and all targets were identified. The total time per scene in these views was approximately 33 minutes.

Scenes 7 and 8, which lasted approximately 335 seconds, comprised a wide and medium view of tank exercises. A field lay between the camera and the tanks, which

conducted exercises next to a tree line approximately 400–600 m away from the camera. Scene 7's camera filmed the scene with the wide view (×1 zoom). Scene 8's camera was zoomed by 2 (twice as many pixels were on the targets). The weather was overcast and drizzly. There was a total of five potential moving targets (tanks moving across the scene in tree line). The videotapes of these scenes were analyzed with the temporal differencing algorithm only. Table 7 summarizes the raw results of the analysis of Scenes 7 and 8. No false alarms occurred in the wide view, and two false alarms occurred in the medium view.

Table 7. Raw Results From the Analysis of Scenes 7 and 8: a Wide and Medium View of Tank Exercises

Target Number	Comments	Wide View Time To Acquire Target (sec)	Medium View Time To Acquire Target (sec)
1	Tank moving R to L	Not acquired	2
2	Tank moving L to R	Not acquired	1
3	Tank moving R to L	5	3
4	Tank moving L to R	Not acquired	0
5	Tank moving R to L	4	4

Scenes 9 and 10 were similar to Scenes 7 and 8. The vista included a field that lay between the camera and tanks conducting exercises near the tree line. Scene 9 was a wide view at about 600 m. Scene 10 was a medium view of the same tank exercises with a ×2 zoom setting. The total tape time was approximately 465 seconds. There was a total of 10 potential moving targets, one of which (target number 7) was a bird flying across the scene. Notably, the bird was not identified as an interesting target by the computer algorithm. Table 8 summarizes the raw results of the analysis of Scenes 9 and 10. There were no false alarms in the wide-view video, and one in the medium-view video.

The last IDA-filmed vista to be analyzed was Scenes 11 and 12. In this vista, a group of three soldiers, relatively well separated, walked along a tree line. A field separated the cameras and the soldiers. Scene 11 was filmed at about a 600-m range, with a ×1 zoom. Scene 12 was filmed at the same range but with a ×2 zoom. In the video of the wide-view scene, distinguishing the soldiers from the trees was quite difficult, even by eye. The scenes were approximately 130 seconds long and contained 3 potential moving targets. Table 9 summarizes the raw results of the analysis of Scenes 11 and 12. Neither scene registered a false alarm.

Table 8. Raw Results From the Analysis of Scenes 9 and 10: a Wide and Medium View of Tank Exercises

Target Number	Comments	Wide View Time To Acquire Target (sec)	Medium View Time To Acquire Target (sec)
1	Tank moving L to R	Not acquired	1
2	Tank moving R to L	Not acquired	Not acquired
3	Tank moving L to R	Not acquired	Not acquired
4	Tank moving R to L	Not acquired	Not acquired
5	Tank moving L to R	Not acquired	0
6	Tank moving R to L	Not acquired	0
7	Bird flying R to L	Not acquired	Not acquired
8	Tank moving L to R	Not acquired	0
9	Tank moving R to L	60	4
10	Tank moving L to R	Not acquired	0

Table 9. Raw Results From the Analysis of Scenes 11 and 12: a Wide and Medium View of Soldier Exercises

Target Number	Comments	Wide View Time To Acquire Target (sec)	Medium View Time To Acquire Target (sec)
1	Humans L to R	Not acquired	8
2	Humans R to L	Not acquired	6
3	Humans L to R	Not acquired	5

CMU provided the results of Scene 13, which were analyzed quantitatively by IDA. This scene varied but primarily contained soldiers or groups of soldiers conducting exercises or walking (either in wooded areas or out in the open). In our analysis, human groups that are not well separated are counted as tracked if the algorithm records and tracks the group. The results of this analysis are different from the previous results because the computer analyzed the scene in real time and the IDA researchers then studied the video-tracking results. Thus, the results are probably closer to what SUO would get in the field with live-time analysis (since results are not distorted by noise in videotape). There was a total of 13 moving targets (or groups of targets) in the approximately 10 minutes of film, primarily shot in a narrow view. The weather varied from sunny to overcast. Table 10 summarizes the raw results of the analysis of Scene 13.

Table 10. Raw Results From the Analysis of Scene 13: a Narrow View of Primarily Soldier Exercises

Target Number	Comments	Time In View (sec)	Time To Acquire (sec)
1	1 Human, R to L	~ 54	0
2	3 Humans, R to L	~ 43	0
3	3 Humans, R to L	~ 23	0
4	2 Humans, L to R	~ 65	5
5	3 Humans, L to R	~ 128	0
6	6 Humans, L to R	~ 30	2
7	1 Human, R to L	~ 7	0
8	3 Humans, R to L	~ 55	0
9	1 Human, R to L,	~ 44	2
10	1 Vehicle, R to L	~ 64	1
11	1 Vehicle, R to L	~ 48	2
12	1 Human, Looming,	~ 28	5
13	2 Humans, Looming	~ 13	7

All targets were acquired in this tape. There were 3 false alarms, 2 of which persisted for approximately 3 seconds each, and 1 that persisted for a fraction of a second. Two of these false alarms appeared to be caused by "track splitting" (i.e., a true target was being tracked and was split into two pieces because of some artifact in the scene). The tracking false alarm is different from a detection false alarm. For instance, if a human target is moving past a vertical object, the track between the human target and the stationary fence pole is split. A false alarm is registered for the fence pole while the human continues to be tracked. This observation is important because the tracking of false alarms is occurring (in this video of real-time analysis) while a real target is being tracked. Thus, this analysis seems to indicate that false alarms will tend to occur when a real target appears in the field of view. In addition, further refinements of this video tracking might mask out such false alarms. The reader should be cautioned, however, that this observation is based on the three false alarms registered in this scene (two of which were tracking false alarms and one of which was a true false alarm) and was not borne out by the previous analyses of the videotapes. That is, in the first sets (Scenes 1–12), false alarms appeared to be caused primarily by the wind. However, since the tape could have been somewhat noisier than live-time video, the false alarms could have been a combination of wind/poor tape quality. Thus, an investigation should be conducted to

determine when false alarms occur during the real-time analysis and would probably show whether split tracks (tracking false alarms) or windy weather (detection false alarms) would be the primary cause of false alarms during real-time analysis.

The last scene (Scene 14), only briefly mentioned here, was a real-time analysis of a busy parking lot at CMU. In this scene, a CCD camera viewed a parking lot in real time. The adaptive background subtraction algorithm was use. The CCD view and real-time analysis were displayed on a monitor. This parking lot included parked cars, moving cars, human groups walking, and single humans walking. The CMU researchers exhibited this experimental analysis to show that they have nearly solved the target ID problem. The previous algorithm had trouble distinguishing human groups from vehicles. This improved algorithm of target recognition appeared to solve the problem by both shape and color analysis. Although individuals in a group cannot be counted, the shape and color algorithm can clearly distinguish groups from vehicles in real time.

The tracking algorithm recorded a total of 67 moving targets in about 15 minutes of watching. Of these 67 targets, approximately 5 were false alarms. Of the 52 cars that moved through the scene, 44 were correctly tracked and identified. Six vehicles were tracked, but misidentified (five times as human and one time as a human group). Two vehicles were never tracked (or identified). In both of these cases, the vehicles were obstructed by another moving vehicle (i.e., two cars moved together through the lot, with one partially obscured from view). The second moving vehicle was always correctly identified. Twenty-nine single humans and two human groups moved through the scene, and all humans and human groups were correctly tracked and identified.

The cause of the false alarms was not clear, although, again, they often appeared to be a result of split tracks (tracking false alarms). No wind was blowing, and the background appeared to be static (no swaying trees). These false alarms did not appear to persist for long times (lasting seconds, at most), but no data are available to quantify this statement.

VII. MEANING OF RESULTS

The raw results in Section VI were analyzed further to address the following questions:

- What is the probability of target detection and how (if at all) does the detection probability depend on the threshold level?
- What is the false-alarm rate, and how does the false-alarm rate depend on the threshold level?
- How long does it take the algorithm to acquire targets?
- Given the measured false-alarm rate, what is the maximum number of false alarms expected as a function of time?

Only the results from the temporal differencing algorithm were used to answer these questions.

Figures 3a–3c show plots of the detection probability and the false-alarm rate versus threshold, and the detection probability versus threshold for the four thresholds examined in detail (see Tables 4 and 5). These data were derived from Scene 3, which was a narrow (close-in) view of soldiers conducting exercises. The detection probability and false-alarm rate fall somewhat for the lowest threshold examined. This appeared to be the case because, with the lower threshold, the algorithm could not detect fast-moving targets. The noisiness of the video could also have contributed to this lower detection rate. Thresholds of 30 and 35 give a 100-percent detection rate but different false-alarm rates. The false-alarm rate is nearly 5 times higher at the lower threshold of 30. Thus, for this analysis, it is clear that a threshold of 35 is the correct threshold to use.

Note that the detection probability (as seen in Figure 3a) is relatively insensitive to the threshold. Changing the threshold only changes the false-alarm rates. The analysis of two views of different scenes with the same threshold does not necessarily mean that they are analyzed with the same sensitivity. In particular, if the views have significantly different zooms (e.g., a range of 600 m with a $\times 1$ and $\times 2$ zoom), the sensitivities at the same threshold value will be different because each camera has a fixed number of pixels at closer-ranged views. In subsequent analyses, a threshold of 35 was used for all

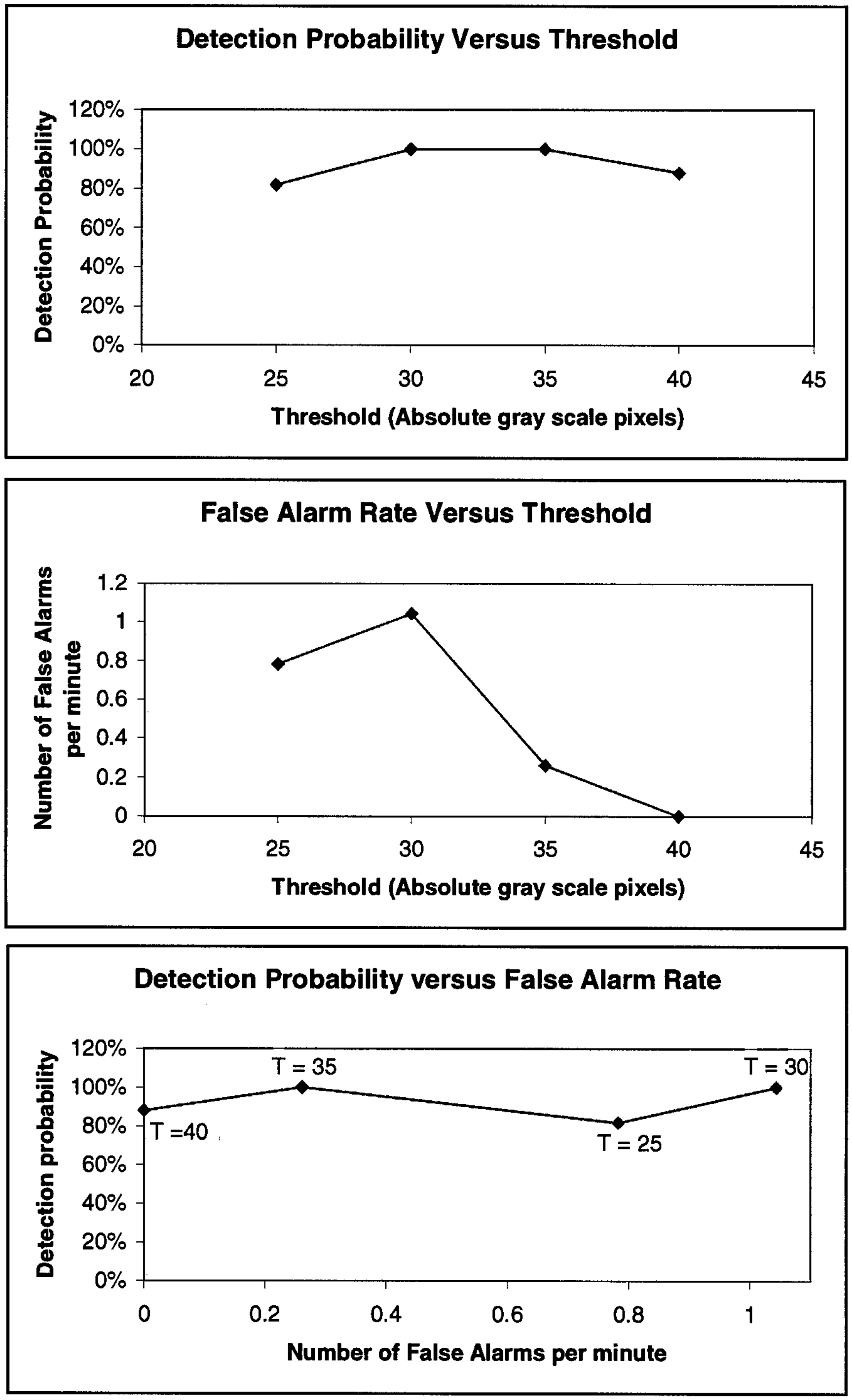

Figure 3a–3c. The Detection Probability and False-Alarm Rate vs. Threshold and the Detection Probability vs. False-Alarm Rate From the Analysis of Scene 3

scenes, regardless of the view (wide, medium, or narrow). The reader should be cautioned that the sensitivities to movements were not identical, because of the different zooms.

Figures 4a–4c show the detection probability vs. the false-alarm rate for all the scenes quantitatively analyzed, using the temporal differencing method of background subtraction. Most analyses had the threshold set to T = 35. The data collected in Scene 13 (from the data provided by CMU) may have been acquired using a different threshold.

The wide-view scenes had the lowest detection rates, never exceeding approximately 58 percent. The narrow and medium views had an average detection rate of 97 percent, with 100-percent detection rate occurring most often. This means that this technique works best when the desired target range is not too far from the camera (approximately 300 m or less). Figure 5 shows the distribution of the number of false alarms per minute recorded per scene for the data shown in Figure 3.

The median of the distribution in Figure 5 is approximately 0.13 false alarms per minute (i.e., half of the recorded false-alarm rates were more than 0.13 per minute and half were less than 0.13 per minute). Many false-alarm rates were 0; however, many of those scenes with a 0 false-alarm rate had modest target detection rates. Table 11 summarizes the results shown in Figures 4 and 5. As mentioned previously, most of the false alarms were caused by the wind.

The most important parameter in determining whether a target is detected is the number of pixels on target. The number of pixels on target is correlated with the view. A medium view (obtained by zooming the camera by 2) is equivalent to 300-m range with no zoom, and a narrow view (camera zoom set to 4) is equivalent to 150-m range with no zoom. Thus, a 2-m moving human at 600 m spans few pixels in an unzoomed camera and is barely detectable (detection probability approximately 39 percent). For a 2-m human who moves closer (at 300 m) or, alternatively, is viewed by a camera with a ×2 zoom setting, the detection probability is approximately 83 percent. The system developed by CMU-Sarnoff researchers includes automatic control of the zoom. This means that one could automatically zoom in on any interesting activity. In addition, by using a system of cameras with varying views and zoom settings, an area can be completely covered out to just about any reasonable range.

This mean number of false alarms per minute (0.22) in the medium and narrow views can be used to estimate the maximum number of false alarms that would be expected during an elapsed period of time. It is assumed that the threshold is set correctly for the scene (at about T = 35 for the gray-scale difference allowed) and that the scene viewed would be a medium-to-narrow view (i.e., 300 m or less for an unzoomed

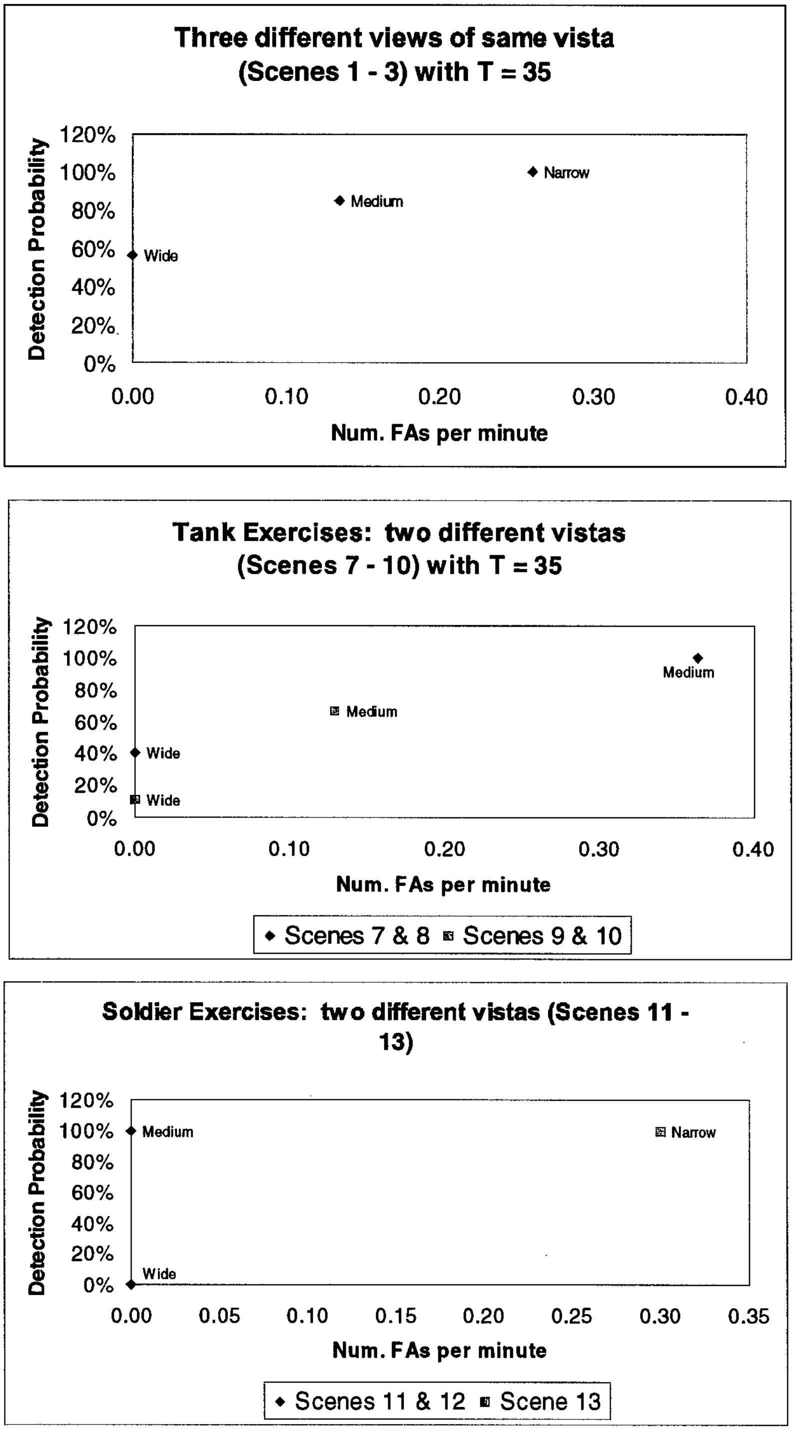

Figure 4a–4c. Detection Probability vs. False-Alarm Rate for Other Scenes Analyzed in This Paper

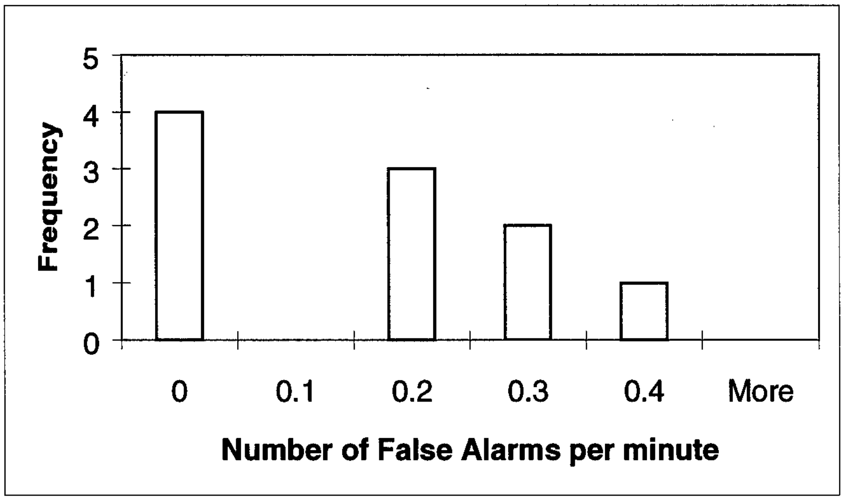

Figure 5. Distribution of the False-Alarm Rates for the Data Shown in Figure 3

Table 11. Summary of Results from Figures 4 and 5

View	Number of Moving Targets	Number of Targets Detected	Number of False Alarms	Total Scene Time (sec)	Probability of Detection (%)	Number of False Alarms Per Minute
Wide	33	13	0	1,375	39	0
Medium	30	25	4	1,372	83	.18
Narrow	30	30	5	1,060	100	0.28
Narrow- and Medium-View totals	60	55	9	2,432	92	0.22

Note for Table 11: *A target was counted as "detected" if the algorithm clearly indicated that it was detected, regardless of the ID assigned.*

camera). Then, even if the conditions are windy, one can expect a mean false-alarm rate of 0.22 per minute. By using Poisson (counting) statistics, we can estimate the maximum number of false alarms expected as a function of the elapsed time, by confidence level. For instance, the Figure 6 shows that there would have been 5 false alarms at the 95-percent confidence level after 20 minutes. This means that in 95 out of 100 experiments (20-minute scenes), the number of false alarms never would have *exceeded* 5 (and would often be fewer than 5 in 20 minutes).

In addition, given the measured dwell time on false alarms (see Table 6), the maximum amount of time spent on false alarms can also be estimated. In the previous

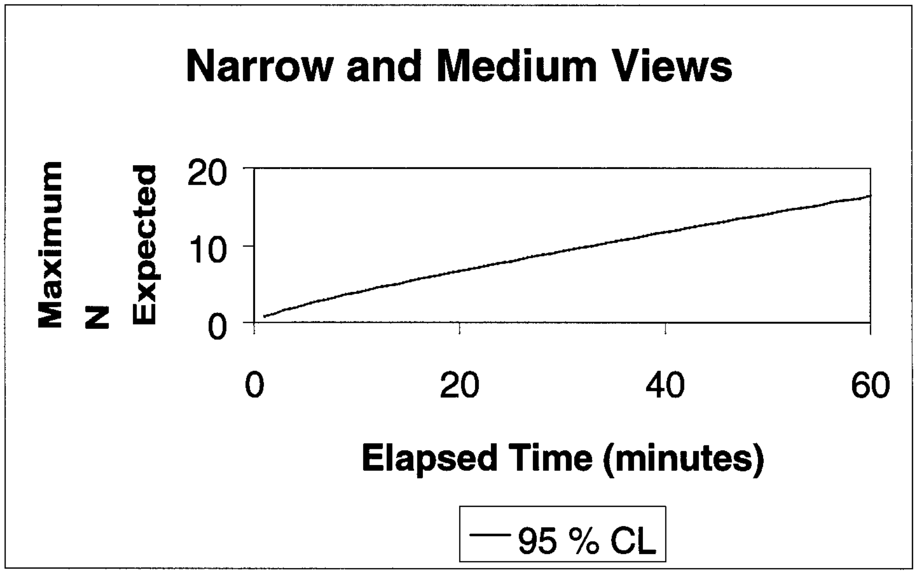

Figure 6. Maximum Total Number of False Alarms Expected as a Function of Elapsed Time (95-percent Confidence Level)

Note for Figure 6: *This was computed by using the mean false-alarm rate and Poisson statistics.*

example, after 20 minutes and no more than 5 false alarms, a total of no more than 1.2 minutes would have been spent on false alarms. That is, although there could be up to 5 false alarms, they occupy less than 6 percent of the total time (see Figure 7).

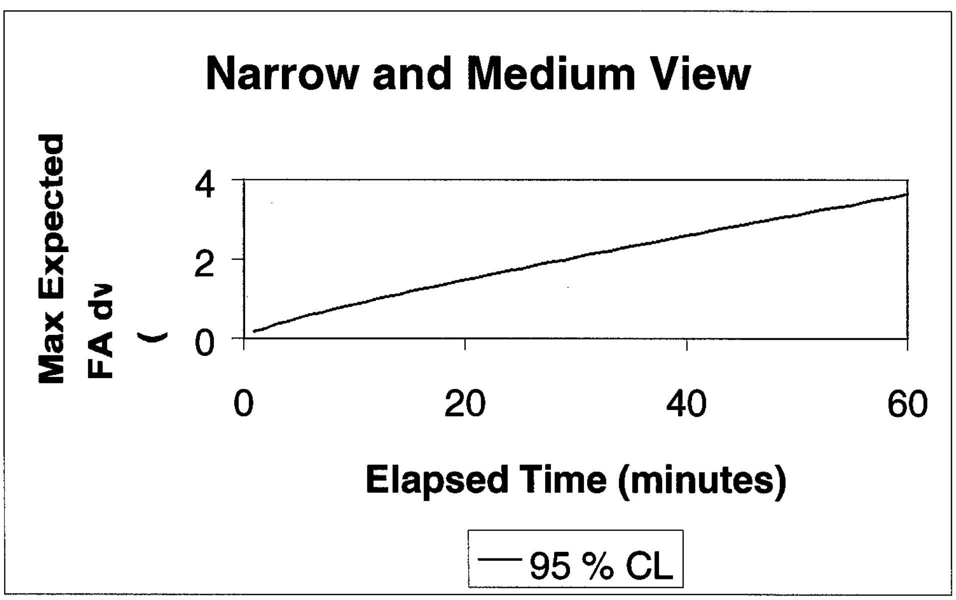

Figure 7. Maximum Total Expected Time Spent on False Alarms as a Function of Elapsed Time (95-percent Confidence Level)

Only the estimates from the narrow and medium view are shown because the wide view had a 0 false-alarm rate.

Lastly, Figures 8a–8c show the estimated times to acquire targets.[4] The first figure shows the time to acquire targets for all data, and the next two figures show the time to acquire the target in the wide view and in the narrow and medium views. More data are present in the narrow- and medium-view plot than in the wide-view plot because more targets were acquired (much higher detection rate) in these higher zoom levels and because the CMU-supplied data (which was a narrow, close-in view) were also included in this plot. In the narrow and medium view, a significant fraction of targets is acquired immediately, and nearly all targets are acquired in less than 10 seconds.

---

4 Targets that were never acquired were omitted from these plots.

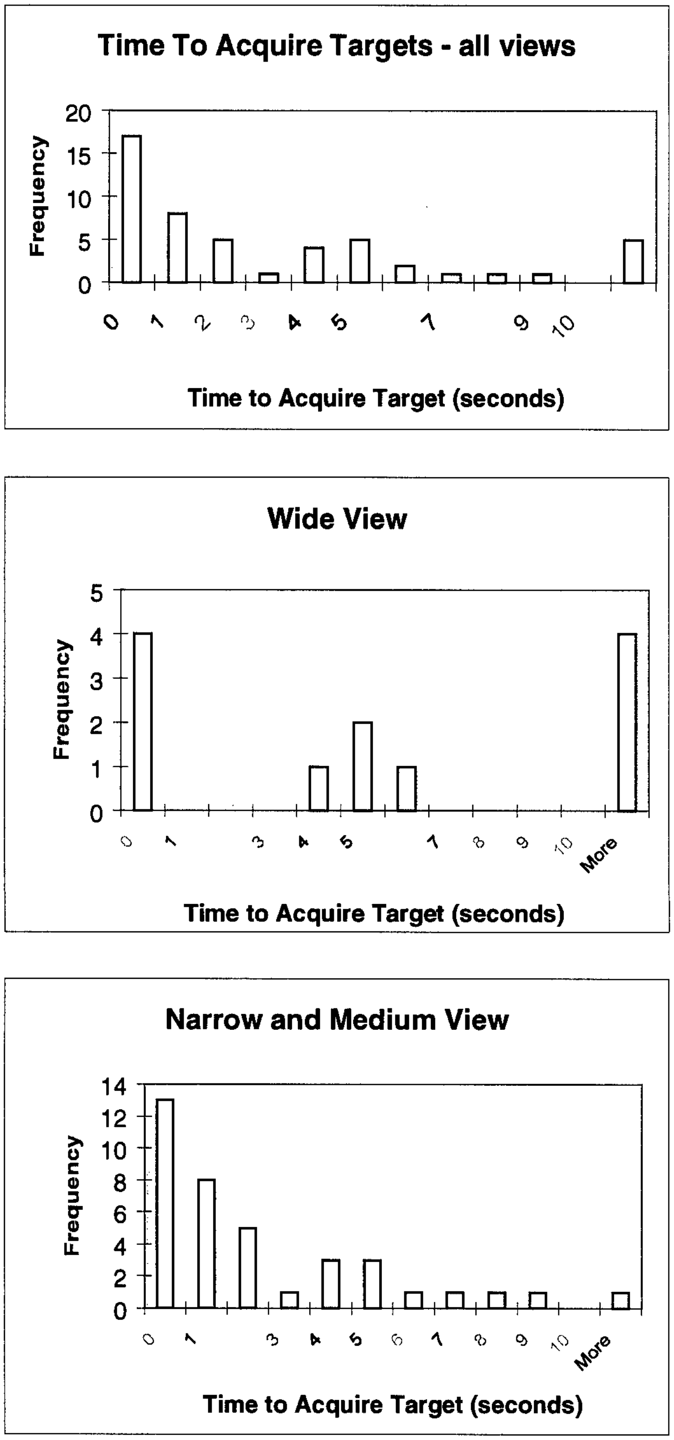

Figure 8a–8c. Time to Acquire Targets for All Data (Top), the Wide-View Data (Middle), and the Closer Ranged Data (Bottom)

VIII. SUMMARY OF FINDINGS

It is clear that the VSAM system is an excellent system for use in automatic detection of targets, particularly when the combination of zoom/range/camera sensitivity gives enough pixels on target. It has a nearly perfect rate of true-target acquisition and fast target-acquisition times and a relatively low rate of false alarms. Furthermore, the false alarms encountered in this analysis might have been caused by poor video quality. With the higher video quality obtained by analyzing live-time images, this false-alarm rate could be significantly lower. In addition, in real time, indications are that the false alarms are caused by tracking false alarms rather than detection false alarms caused by background artifacts.

Preliminary results from new algorithms that discriminate targets better are encouraging. The CMU researchers will probably solve this problem in the near future.

Commercial parts, such as CCD cameras and PCs, are used to build these systems. The hardware is easy to obtain, integrate, and use and, most importantly, is inexpensive. Most of the system costs are the up-front costs of developing the software and tracking algorithms. Thus, most of the cost is developmental—not manufacturing. We estimate that one CCD/PC unit could ultimately cost less than $7,500.

This system, however, is being developed for battlefield management, not SUO applications. Thus, the developers would have to adapt the system for SUO. Researchers believe that the systems need to be "ruggedized" for a variety of field situations. In addition, developers would have to consider a few critical areas:

- **Power and weight.** The camera/PC system needs to be made lightweight and battery powered so that small units can carry and operate it in the field.
- **Communications.** Developers need to address the communications and bandwidth problems that are peculiar to SUO.
- **Integration.** The integration issues of a system of cameras for SUO may be different from what is being developed for the battlefield management project.

The power and weight issues can probably be addressed with adaptations of commercial products (e.g., using laptop-style computers and providing lithium batteries to

power the CCD cameras). It may be possible to develop specialized digital electronics [application-specific integrated circuits (ASICs) or digital signal processors (DSPs)] to handle much of the detection and tracking, thus further reducing the power requirements. The researchers are already addressing communications in their present program. Many of the ideas overlap, but some would have to be recast for the SUO program. Lastly, the integration of multiple tracking systems into a single system useful to SUO may be different because it would probably involve a different number of systems (perhaps more densely packed). The CMU-Sarnoff researchers would be able to address the communications and integration issues, and other development projects within SUO could address the remaining issues, such as power and weight.

As noted in the introduction, VSAM, as conceived, is a system of coordinated sensors. However, the data presented in this paper were from a single sensor. Several coordinated sensors could view an area of interest with several different zooms and at slightly different angles, which would reduce the false-alarm rate substantially. For example, if each sensor is recording a false-alarm rate of 0.22 per minute, the two coordinated cameras viewing the same scene would have a false-alarm rate as low as 0.05 per minute. Such a system could be developed and optimized, taking into account the numbers of sensors desired, the range of detection, the quality of the sensors, the maximum false-alarm rate allowed, and the types of scenes examined. This analysis shows that even a single sensor can detect motion over a relatively long range and with a fairly low rate of false alarms. A system of such sensors using the research conducted by the CMU-Sarnoff team as a basis would be a potentially powerful tool for SUO monitoring of activity.

In summary, the results of this analysis show that the VSAM technology, although still under development, demonstrated unique capabilities that SUO should adapt and pursue for its own applications. The CMU researchers and Sarnoff Corporation have developed algorithms that show great promise for SUO applications.

GLOSSARY

APC	Armored Personnel Carrier
ASIC	application-specific integrated circuit
BAA	Broad Agency Announcement
CCD	charged-coupled device
CMU	Carnegie Mellon University
COTS	commercial off-the-shelf
DARPA	Defense Advanced Research Projects Agency
DSP	digital signal processor
ID	identification
IDA	Institute for Defense Analyses
IR	infrared
m	meters
PC	personal computer
R&D	research and development
SUO	Small Unit Operations
SUO/SAS	Small Unit Operations/Situational Awareness System
VSAM	Video Surveillance and Monitoring

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
December 1999

3. REPORT TYPE AND DATES COVERED
Final—August 1998–November 1999

4. TITLE AND SUBTITLE
Analysis of VSAM Research and Carnegie Mellon University and the Sarnoff Corporation: Potential Application to Small Unit Operations

5. FUNDING NUMBERS
DASW01 98 C 0067
DA-2-210

6. AUTHOR(S)
Cynthia Dion-Schwarz

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311-1772

8. PERFORMING ORGANIZATION REPORT NUMBER
IDA Paper P-3428

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency/TTO
3701 N. Fairfax Drive
Arlington, VA 22201

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; unlimited distribution (2/28/2000).

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 180 words)*

This paper summarizes an analysis of the Video Surveillance andMonitoring (VSAM) research being conducted by researchers at Carnegie Mellon University and the Sarnoff Corporation under the Defense Advanced Research Projects Agency (DARPA) Image Understanding Program. The reachers' goal is to develop a cooperative, multi-sensor video surveillance system for large battfield areas. The team is developing software and integrating inexpensive commercial-off-the-shelf (COTS) hardware systems that will automatically track and identify moving targets. The team is concentrating on integrating the hardware and developing the coordinated tracking algorthims (software) that provide good target identification (ID) and target tracking with a low rate of false alarms. Although the researchers are stydying ways to create a VSAM for battlefield management, this system also has clear applications to Small Unit Operations (SUO).

14. SUBJECT TERMS
Defense Advanced Research Projects Agency (DARPA), false-alarm rate, motion detection, Situational Awareness System (SAS), Small Unit Operations (SUO), target identification (ID), target tracking, Video Surveillance and Monitoring (VSAM)

15. NUMBER OF PAGES
37

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102